

Suicide ideation of individuals in online social networks

Naoki Masuda^{1*}, Issei Kurahashi², Hiroko Onari²

¹ Department of Mathematical Informatics,
The University of Tokyo,
7-3-1 Hongo, Bunkyo, Tokyo 113-8656, Japan

² iAnalysis LLC,
2-2-15 Minamiaoyama, Minato-ku, Tokyo 107-0062, Japan

* Corresponding author
Email: masuda@mist.i.u-tokyo.ac.jp Tel: +81-3-5841-6931

August 21, 2012

Abstract

Suicide is a major cause of death for adolescents in many countries. The impact of social isolation on suicide in the context of explicit social networks of individuals is relatively unexplored. We statistically examined relationships between suicide ideation and user's characteristics using a large data set obtained from a major social networking service in Japan. We found that the number of user-defined communities to which a user belongs to, the intransitivity (i.e., paucity of triangles including the user), and the fraction of suicidal neighbors in the social network, contributed the most to suicide ideation in this or-

der. Age and gender contributed little. We also found similar results for depressive symptoms.

Keywords: suicide, depression, clustering, transitivity, social isolation, social networking service

1 Introduction

Suicide is a major cause of death in many countries. Japan possesses the highest suicide rate among the OECD countries in 2009 (Chambers, 2010). In fact, suicide explains the largest number of death cases for Japanese adolescents in their twenties and thirties (Chambers, 2010). Suicide is also a major cause of death for youths in other countries including the United States (of the Census, 2012).

Since the seminal sociological study by Durkheim in the late nineteenth century (Durkheim, 1951), suicides have been studied for both sociology interests and public health reasons. In particular, Durkheim and later scholars pointed out that social isolation, also referred to as the lack of social integration, is a significant contributor to suicidal behavior (Durkheim, 1951; Trout, 1980; Joiner Jr. *et al.*, 2005; Wray *et al.*, 2011). Roles of social isolation in inducing other physical and mental illnesses have also been examined (Putnam, 2001). Conceptual models that inherit Durkheim’s idea also claim that social networks affect general health conditions including tendency to suicide (Bearman, 1991; Berkman *et al.*, 2000; Kawachi & Berkman, 2001).

Social network analysis provides a pragmatic method to quantify social isolation (Wasserman & Faust, 1994; Newman, 2010). In their seminal work, Bearman and Moody explicitly studied the relationship between suicidal behavior and egocentric social networks for American adolescents using data obtained from a national survey (National Longitudinal Study of Adolescent Health) (Bearman & Moody, 2004). They showed that, among many independent variables including those unrelated to social networks, a small number of friends and a small fraction of triangles to which an individual belongs significantly contribute to suicide ideation and attempts. A small number of friends is an intuitive indicator of social isolation. Another study derived from self reports from Chinese adolescents also supports this idea in a quantitative manner (Cui *et al.*, 2010). The paucity of triangles, or intransitivity (Wasserman & Faust, 1994), also characterizes social isolation (Bearman & Moody, 2004). Individuals without triangles are

considered to lack membership to social groups even if they have many friends (Krackhardt, 1999); social groups are often approximated by overlapping triangles (Palla *et al.*, 2005; Onnela *et al.*, 2007).

Nevertheless, the structure of the Bearman–Moody study (Bearman & Moody, 2004) implies that our understanding of relationships between social networks and suicide is still limited. First, in the survey, a respondent was allowed to list best five friends of each gender. However, many respondents would generally have more friends. The imposed upper limit may distort network-related personal quantities such as the number of friends and triangles. Second, their study was confined inside each school in the sense that only in-school names are matched. If a respondent X named two out-school friends that are actually friends of each other, the triangle composed of these three individuals was dismissed from their statistical analysis. Therefore, the accuracy of the triangle counts in their study may be limited such that the relationship between intransitivity and suicidal behavior remains elusive.

In the present study, we examine the relationship between social networks and suicide ideation using a data set obtained from a dominant social networking service (SNS) in Japan, named mixi. Our approach addresses limitations in the previous study (Bearman & Moody, 2004). First, an entire social network of users is available, where a link between two users represents explicit bidirectional friendship that both users have endorsed. Some users have quite a large number of friends, as in general social networks (Newman, 2010). Second, for the same reason, the number of triangles for each user is calculated without error. An additional feature of the present data set is that the population is relatively diverse because anybody can register for free. In contrast, the respondents are 7 to 12 graders in schools in the Bearman–Moody study.

A function of mixi relevant to this study is user-defined communities. A community is a group of users that get together under a common interest, such as hobby, affiliation, and creed. A user-defined community of mixi is often composed of users that have not known each other beforehand. Although some SNSs have user-defined communities, and their dynamics were studied (Backstrom *et al.*, 2006), major SNSs including Facebook do not own this type of user-defined communities. We define suicide ideation by the membership of a user to at least one community related to suicide. Then, we statistically compare users with and without suicide ideation in terms of users’ properties including those related to egocentric networks.

2 Methods

2.1 Data sets

Mixi is a major SNS in Japan. It started to operate on March 2004 and enjoys more than 2.7×10^7 registered users as of March 2012. Similar to other known SNSs, users of mixi can participate in various activities such as making friendship with other users, tweeting, sending instant messages to others, uploading photos, and playing online games. Registration is free.

In mixi, there are more than 4.5×10^6 user-defined communities on various topics as of April 2012. Users can join a user-defined community if the owner personally permits or allows anybody to join.

We identify suicide ideation with the membership of a user to at least one suicidal community. To define suicidal community, which is sufficiently active, we first select communities satisfying the following five criteria: (1) The name includes the word “suicide” (“jisatsu” in Japanese), (2) there are at least 1000 members on November 2, 2011, (3) there are at least 100 comments posted on October, 2011, which are directed to other comments or topics, (4) there are at least three independent topics on which comments were made on October, 2011, and (5) the condition for admission is made open to public. Seven communities meet these criteria. Then, we excluded one community whose name indicates that it concentrates on methodologies of committing suicide and two communities whose names indicate that they encourage members to live with hopes (one contains the word “want to live”, and the other contains the word “have a fun” in their names; translations by the authors).

As a result, four communities are qualified as suicidal communities. The user statistics of these communities are shown in Tab. 1. A user that belongs to at least one suicidal community is defined to possess suicide ideation. To exclude inactive users, we restricted ourselves to the set of active users. The active user is defined as users that existed as of January 23, 2012 and logged on to mixi in more than 20 days per month on average from August through December 2011. A similar definition was used in a previous study of the Facebook social network (Ugander *et al.*, 2011). We also discarded users with zero or one friend on mixi because the triangle count that we describe below is undefined for such users. Despite this exclusion, the remaining data allow us to examine the effect of social isolation in terms of the degree, i.e., number of neighbors, because the degree is widely distributed between 2 and 1000. There are 9990 active users with suicide ideation (suicide group).

We statistically compare the users in the suicide group with users without suicide ideation. Because the number of users is huge, we randomly select

228949 active users that possess at least two friends and belong to neither of the seven candidates of the suicidal community defined above nor the ten candidates of the depression-related community (see Appendix for the analysis of depressive symptoms). We call this set of users the control group.

The employees of mixi deleted private information irrelevant to the present study and encrypted the relevant private information before we analyzed the data. In addition, we conducted all the analysis in the central office of mixi located in Tokyo using a computer that is not connected to Internet.

2.2 Statistical models

The dependent variable that represents the level of suicide ideation is binary, i.e., whether a user belongs to a suicidal community or not. Therefore, we use univariate and multivariate logistic regressions. To check the multicollinearity between independent variables to justify the use of the multivariate logistic regression, we carry out two subsidiary analysis. First, we measure the variance inflation factor (VIF) for each independent variable (see (Stine, 1995; Tufféry, 2011) and references therein). The VIF is the reciprocal of the fraction of the variance of the independent variable that is not explained by linear combinations of the other independent variables. It is recommended that the VIF value for each independent variable is smaller than 10 (preferably smaller than 5) for the multivariate logistic regression to be valid. Second, we measure the Pearson, Spearman, and Kendall correlation coefficients between the independent variables.

To quantify the explanatory power of the logistic model, we measure the area under the receiver operating characteristic curve (AUC) for each fit (e.g., (Tufféry, 2011)). The receiver operating characteristic curve is the trajectory of the false positive (i.e., fraction of users in the control group that are mistakenly classified into the suicide group on the basis of the linear combination of the independent variables) and the true positive (i.e., fraction of users in the suicide group correctly classified into the suicide group), when the threshold for classification is varied. The AUC value falls between 0.5 and 1. A large AUC value indicates that the logistic regression fits well to the data in the sense that users are accurately classified into suicide and control groups.

2.3 Independent Variables

We consider seven independent variables. Their univariate statistics for the suicide and control groups are shown in Tab. 2.

Demographics. Demographic independent variables include age and gender. Our analysis does not include ethnic components because most users of mixi are Japanese-speaking Japanese; mixi provides services in Japanese. Other demographic, socioeconomic, and personal characteristic variables such as residence area, occupation, company/school, and hobby, are not used because they are unreliable; many users leave them blank or do not fill them consistently, probably because they do not want to disclose them.

Community number. The number of user-defined communities that a user belongs to is selected as an independent variable. We refer to this quantity as community number. The community number obeys a long tailed distribution for both suicide and control groups (Fig. 1). The mean is quite different between the two groups (Tab. 2).

Degree. When a user sends a request to another user and the recipient accepts the request, the pair of users form an undirected social tie, called Friends. A web of Friends defines a social network of mixi. We adopt degree as the most basic network-related independent variable. The degree is the number of neighbors (i.e., Friends), and denoted by k_i for user i . The system of mixi allows a user to own at most degree 1000. As is consistent with the previous analysis of a much smaller data set of mixi (Yuta *et al.*, 2007), the degree distributions for both groups are long tailed (Fig. 2). A small degree is an indicator of social isolation.

Local clustering coefficient We quantify transitivity, or the density of triangles around a user, by the local clustering coefficient, denoted by C_i for user i . A directed-link version of the same quantity is used in the Bearman–Moody study. For user i having degree k_i , there can be maximum $k_i(k_i - 1)/2$ triangles that include user i . We define C_i as the actual number of triangles that include i divided by $k_i(k_i - 1)/2$. Some examples are shown in Fig. 3. By definition, $0 \leq C_i \leq 1$. We discarded the users with $k_i \leq 1$ because C_i is defined only for users with $k_i \geq 2$.

C_i quantifies the extent to which neighbors of user i are adjacent to each other (Watts & Strogatz, 1998; Newman, 2010). If C_i is large, the user is considered to be embedded in close-knit social groups (Wasserman & Faust, 1994; Watts & Strogatz, 1998; Newman, 2010). A small C_i value is an indicator of social isolation.

As in many networks (Newman, 2010), C_i decreases with k_i in both suicide and control groups (Fig. 4). Therefore, we will carefully distinguish the influence of k_i and C_i on suicide ideation by combining univariate and multivariate regressions.

Homophily. Suicide may be a contagious phenomenon (e.g., (Mann, 2002; Baller & Richardson, 2002; Romer *et al.*, 2006; Hedström *et al.*, 2008; Baller & Richardson, 2009; Wray *et al.*, 2011)). If so, a user is inclined to

suicide ideation when a neighbor in the social network does. Therefore, we adopt the fraction of neighbors with suicide ideation as an independent variable. It should be noted that, even if a user with suicide ideation has relatively many friends with suicide ideation, it does not necessarily imply that suicide is contagious. Homophily may be a cause of such assortativity. In this study, we do not attempt to differentiate the effect of imitation and homophily. The differentiation would require analysis of temporal data (Aral *et al.*, 2009; Shalizi & Thomas, 2011). Nevertheless, for a notational reason, we refer to the fraction of neighbors as the homophily variable.

Registration period. A user that registered to mixi long time ago may be more active and own more resources in mixi than new users. Such an experienced user may tend to simultaneously have, for example, a large community number, large degree, and perhaps high activities in various communities including suicidal ones. To control for this factor, we measure the registration period defined as the number of days between the registration date and January 23, 2012.

3 Results

Table 2 indicates that the difference in the mean of each independent variable between the suicide and control groups is significant ($p < 0.001$, Student’s t -test). We also verified that the distributions of each independent variable are also significantly different between the two groups ($p < 0.0001$, Kolmogorov-Smirnov test).

The results obtained from the multivariate logistic regression are summarized in Tab. 3. The VIF values (see Methods) are much less than 5 for all the independent variables. The three types of correlation coefficients between pairs of the independent variables are also sufficiently small (Tab. 4). On these bases, we justify the application of the multivariate logistic regression to our data.

The odds ratio (OR) values shown in Table 3 suggest the following. A one-year older user is 1.00463 times more likely to belong to the suicide group than the control group on average. Likewise, being female, membership to one community, having one friend, an increase in C_i by 0.01, an increase in the fraction of friends in the suicide group (i.e., homophily variable) by 0.01, and one day of the registration period make a user 0.821, 1.00733, 0.99790, $0.0093^{0.01} = 0.95$, $(2.22 \times 10^{12})^{0.01} = 1.33$, and 0.999383 times more likely to belong to the suicide group, respectively. For all the independent variables, the 95 % confidence intervals of the ORs do not contain unity, and

the p -values are small. Therefore, all the independent variables significantly contribute to the regression. In addition, because the AUC (see Methods) is large (i.e. 0.873), the estimated multivariate logistic model captures much of the variation in the user’s behavior, i.e., whether to belong to the suicide group or not.

All the independent variables significantly contribute to the multivariate regression probably because of the large sample size of our data set. Therefore, we carried out the univariate logistic regression between the dependent variable (i.e., membership to the suicide versus control group) and each independent variable to better clarify the contribution of each independent variable.

The results of the univariate logistic regression are shown in Tab. 5. Although the p -value for each independent variable is small, the AUC value considerably differs by the independent variable.

The ORs for the community number, local clustering coefficient, homophily, and registration period are consistent between the multivariate and univariate regressions. For example, both regressions indicate that a user with a large community number tends to belong to the suicide group. These independent variables also yield large AUC values under the univariate regression.

The community number makes by far the largest contribution among the seven independent variables. The AUC value obtained from the univariate regression (0.867) is close to that obtained by the multivariate regression (0.873).

The independent variable with the second largest explanatory power is the local clustering coefficient (AUC = 0.690). The results are consistent with the previous ones (Bearman & Moody, 2004). We stress that we reach this conclusion on the basis of the data set whose full social network is available.

The homophily variable makes the third largest contribution (AUC = 0.643). Although we refer to this independent variable as homophily (see Methods), the effect of this variable is in fact interpreted as either homophily or contagion (Aral *et al.*, 2009; Shalizi & Thomas, 2011). Nevertheless, the result is consistent with previous claims that suicide is contagious (for recent accounts, see (Mann, 2002; Baller & Richardson, 2002; Romer *et al.*, 2006; Hedström *et al.*, 2008; Baller & Richardson, 2009; Wray *et al.*, 2011)) and that other related states such as depressive symptoms are contagious (Christakis & Fowler, 2009; Rosenquist *et al.*, 2011) (but see (Lyons, 2011; VanderWeele, 2011)).

The effect of the age, gender, and degree (i.e., number of friends), on suicide ideation is small, yielding small AUC values, close to the minimum value 0.5 (Tab. 5). In addition, the ORs for these variables are inconsistent

between the multivariate and univariate regressions. For example, a female user is more likely to belong to the suicide group according to the univariate regression and vice versa according to the multivariate regression. Therefore, we conclude that these three independent variables do not explain suicide ideation.

The registration period also yields a small AUC value (i.e., 0.545). Therefore, dependence of suicide ideation on the other independent variables is not derived from common dependency of these variables on the registration period.

Our data set allows us to investigate correlates between users' other characteristics and the independent variables if the characteristics have corresponding used-defined communities in mixi. We repeated the same series of analysis for depressive symptoms, which are suggested to be implicated in suicidal behavior (Mann, 2002; Joiner Jr. *et al.*, 2005; Brezo *et al.*, 2006). A user is defined to own depressive symptoms when the user belongs to at least one of the seven depression-related communities (Appendix). The results of the statistical analysis are similar to those for suicide ideation (Appendix).

4 Discussion

We investigated relationships between suicide ideation and personal characteristics including social network variables using the data obtained from a major SNS in Japan. We found that an increase in the community number (i.e., the number of user-defined communities to which a user belongs), decrease in the local clustering coefficient (i.e., local density of triangles, or transitivity), and increase in the homophily variable (i.e., fraction of neighboring users with suicide ideation) contribute to suicide ideation by the largest amounts in this order. In addition, the results are qualitatively the same when we replaced suicide ideation by depressive symptoms. Remarkably, the most significant three variables represent online social behavior of users rather than demographic properties such as the age and gender.

Our result that the age and gender little influence suicide ideation is inconsistent with previous findings (Wray *et al.*, 2011). The weak age effect in our result may be because the majority of registered users is young; the mean age of the users in the control group is 27.7 years old (Tab. 2). Nevertheless, we stress that suicide is a problem particularly among young generations to which a majority of the users belong.

Our result that the degree little explains suicide ideation is inconsistent with previous studies that explicitly examined the effect of the number friends in social networks on suicide (Bearman & Moody, 2004; Cui *et al.*, 2010)

and with the long-standing claim that social isolation elicits suicidal behavior (Durkheim, 1951; Trout, 1980; Joiner Jr. *et al.*, 2005; Wray *et al.*, 2011). As compared to typical users, some users may spend a lot of time online to gain many ties with other users and belong to many communities on the SNS. Nevertheless, such a user may be active exclusively online and feel lonely, for example, to be prone to suicide ideation. Although this is a mere conjecture, such a mechanism would also explain the strong contribution of the community number to suicide ideation revealed in our analysis.

We used a data set representing user behavior online. Nowadays, many people, especially the young, regularly devote much time to online activities including SNSs (Martin, 2010). Therefore, the data obtained from SNSs are considered to capture a significant part of users' lives.

Because mixi enjoys a huge number of users and implements the user-defined community as a main function, user-defined communities of mixi cover virtually all major topics. Therefore, applying the present methods to other psychiatric illness and symptoms, such as schizophrenia, bipolar disorder, and alcohol abuse, as well as positive symptoms are expected to be profitable.

Acknowledgments

We thank mixi, Inc. for providing us with their data and Taro Takaguchi for careful reading of the manuscript. We acknowledge financial supports provided through Grants-in-Aid for Scientific Research (No. 23681033).

References

- [Aral *et al.*, 2009] Aral, S., Muchnik, L. & Sundararajan, A. 2009. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proc. Natl. Acad. Sci. USA*, 106, 21544–21549.
- [Backstrom *et al.*, 2006] Backstrom, L., Huttenlocher, D., Kleinberg, J. & Lan, X. 2006. Group formation in large social networks: membership, growth, and evolution. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 44–54.
- [Baller & Richardson, 2002] Baller, R. D. & Richardson, K. K. 2002. Social integration, imitation, and the geographic patterning of suicide. *Amer. Soc. Rev.* 67, 873–888.

- [Baller & Richardson, 2009] Baller, R. D. & Richardson, K. K. 2009. The “dark side” of the strength of weak ties: the diffusion of suicidal thoughts. *J. Health and Soc. Behav.* 50, 261–276.
- [Bearman, 1991] Bearman, P. S. 1991. The social structure of suicide. *Sociol. Forum*, 6, 501–524.
- [Bearman & Moody, 2004] Bearman, P. S. & Moody, J. 2004. Suicide and friendships among American adolescents. *American Journal of Public Health*, 94 (1), 89–95.
- [Berkman *et al.*, 2000] Berkman, L. F., Glass, T., Brissette, I. & Seeman, T. E. 2000. From social integration to health: Durkheim in the new millennium. *Soc. Sci. Med.* 51, 843–857.
- [Brezo *et al.*, 2006] Brezo, J., Paris, J. & Turecki, G. 2006. Personality traits as correlates of suicidal ideation, suicide attempts, and suicide completions: a systematic review. *Acta Psychiatrica Scandinavica*, 113, 180–206.
- [Chambers, 2010] Chambers, A. 2010. Japan: ending the culture of the ‘honourable’ suicide. *The Guardian*, (3 August 2010).
- [Christakis & Fowler, 2009] Christakis, N. A. & Fowler, J. H. 2009. *Connected*. Little, Brown and Company, New York.
- [Cui *et al.*, 2010] Cui, S., Cheng, Y., Xu, Z., Chen, D. & Wang, Y. 2010. Peer relationships and suicide ideation and attempts among Chinese adolescents. *Child Care Health and Development*, 37, 692–702.
- [Durkheim, 1951] Durkheim, E. 1951. *Suicide*. Free Press, New York.
- [Hedström *et al.*, 2008] Hedström, P., Liu, K. & Nordvik, M. 2008. Interaction domains and suicide: a population-based panel study of suicides in Stockholm, 1991–1999. *Soc. Forces*, 87 (2), 713–740.
- [Joiner Jr. *et al.*, 2005] Joiner Jr., T. E., Brown, J. S. & Wingate, L. R. 2005. The psychology and neurobiology of suicidal behavior. *Annu. Rev. Psychol.* 56, 287–314.
- [Kawachi & Berkman, 2001] Kawachi, I. & Berkman, L. F. 2001. Social ties and mental health. *J. Urban Health*, 78, 458–467.
- [Krackhardt, 1999] Krackhardt, D. 1999. The ties that torture: Simmelian tie analysis in organizations. *Research in the Sociology of Organizations*, 16, 183–210.

- [Lyons, 2011] Lyons, R. 2011. The spread of evidence-poor medicine via flawed social-network analysis. *Statistics, Politics, and Policy*, 2 (1), Article 2.
- [Mann, 2002] Mann, J. J. 2002. A current perspective of suicide and attempted suicide. *Ann. Internal Med.* 136, 302–311.
- [Martin, 2010] Martin, D. 2010. What Americans do online: social media and games dominate activity. *Nielsen News, Online*, (2 August 2010).
- [Newman, 2010] Newman, M. E. J. 2010. *Networks — An introduction*. Oxford University Press, Oxford.
- [of the Census, 2012] of the Census, U. B. 2012. Statistical abstract of the United States.
- [Onnela *et al.*, 2007] Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J. & Barabási, A. L. 2007. Structure and tie strengths in mobile communication networks. *Proc. Natl. Acad. Sci. USA*, 104, 7332–7336.
- [Palla *et al.*, 2005] Palla, G., Derényi, I., Farkas, I. & Vicsek, T. 2005. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435, 814–818.
- [Putnam, 2001] Putnam, R. D. 2001. *Bowling Alone*. Simon & Schuster.
- [Romer *et al.*, 2006] Romer, D., Jamieson, P. E. & Jamieson, K. H. 2006. Are news reports of suicide contagious? A stringent test in six U. S. cities. *J. Communication*, 56, 253–270.
- [Rosenquist *et al.*, 2011] Rosenquist, J. N., Fowler, J. H. & Christakis, N. A. 2011. Social network determinants of depression. *Mol. Psychiatry*, 16, 273–281.
- [Shalizi & Thomas, 2011] Shalizi, C. R. & Thomas, A. C. 2011. Homophily and contagion are generically confounded in observational social network studies. *Sociol. Methods Res.* 40, 211–239.
- [Stine, 1995] Stine, R. A. 1995. Graphical interpretation of variance inflation factors. *Am. Stat.* 49, 53–56.
- [Trout, 1980] Trout, D. L. 1980. The role of social isolation in suicide. *Suicide and Life-Threatening Behav.* 10, 10–23.

- [Tufféry, 2011] Tufféry, S. 2011. *Data Mining and Statistics for Decision Making (2nd edition)*. Willey, Chichester.
- [Ugander *et al.*, 2011] Ugander, J., Karrer, B., Backstrom, L. & Marlow, C. 2011. The anatomy of the Facebook social graph. arXiv:1111.4503, .
- [VanderWeele, 2011] VanderWeele, T. J. 2011. Sensitivity analysis for contagion effects in social networks. *Sociol. Methods & Res.* 40, 240–255.
- [Wasserman & Faust, 1994] Wasserman, S. & Faust, K. 1994. *Social Network Analysis*. Cambridge University Press, New York.
- [Watts & Strogatz, 1998] Watts, D. J. & Strogatz, S. H. 1998. Collective dynamics of ‘small-world’ networks. *Nature*, 393, 440–442.
- [Wray *et al.*, 2011] Wray, M., Colen, C. & Pescosolido, B. 2011. The sociology of suicide. *Annu. Rev. Sociol.* 37, 505–528.
- [Yuta *et al.*, 2007] Yuta, K., Ono, N. & Fujiwara, Y. 2007. A gap in the community-size distribution of a large-scale social networking site. arXiv:0708.3063, .

Figure captions

Figure 1: Distribution of the community number (i.e., number of communities to which a user belongs) for the suicide, depression, and control groups. We set the bin width for generating the histogram to 50. The abrupt increase in the distribution at 1000 communities for the suicide and depression groups is owing to the restriction that a user can belong to at most 1000 communities.

Figure 2: Complementary cumulative distribution of the degree (i.e., fraction of users having the degree larger than a specified value) for the suicide, depression, and control groups.

Figure 3: Examples of the degree (k_i) and the local clustering coefficient (C_i). The shown values of k_i and C_i are for the black nodes.

Figure 4: Dependence of the mean local clustering coefficient on the degree for the suicide, depression, and control groups. Each data point $C(k)$ for degree k is obtained by averaging C_i over the users in a group with degree k . Large fluctuations of $C(k)$ at large k values are caused by the paucity of users having large k .

Table captions

Table 1: Statistics of suicidal communities.

Table 2: Univariate statistics of independent variables for the suicide and control groups. The p -value for the gender is based on the Chi-square test. The p -values for the other independent variables are based on the Student's t -test. Also shown are the statistics of two auxiliary variables that are not used in the logistic regression, i.e., the number of suicidal communities to which the user belongs and the number of days on which the user logged on to mixi. The p -value for the number of log-on days is based on the Student's t -test. SD: standard deviation.

Table 3: Multivariate logistic regression of suicide ideation on individual and network variables. OR: odds ratio; CI: 95 % confidence interval; VIF: variance inflation factor.

Table 4: Correlation coefficients between pairs of independent variables for the suicide, depression, and control groups. P: Pearson; S: Spearman; K: Kendall correlation coefficients.

Table 5: Univariate logistic regression of suicide ideation on individual and network variables. OR: odds ratio; CI: 95 % confidence interval; AUC: area under the curve.

Figure 1

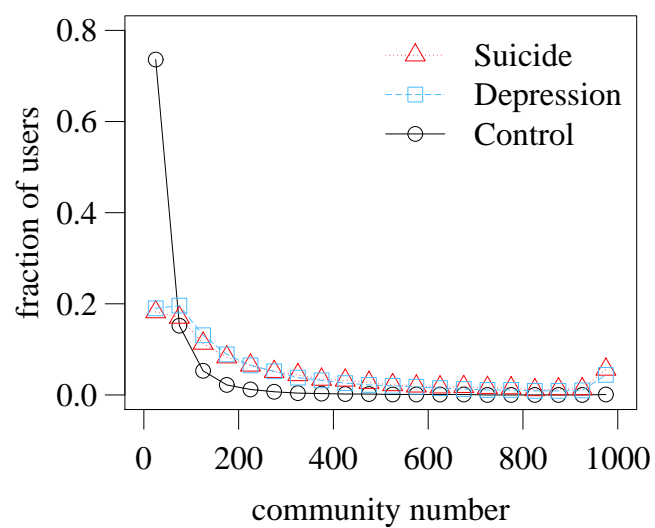


Figure 2

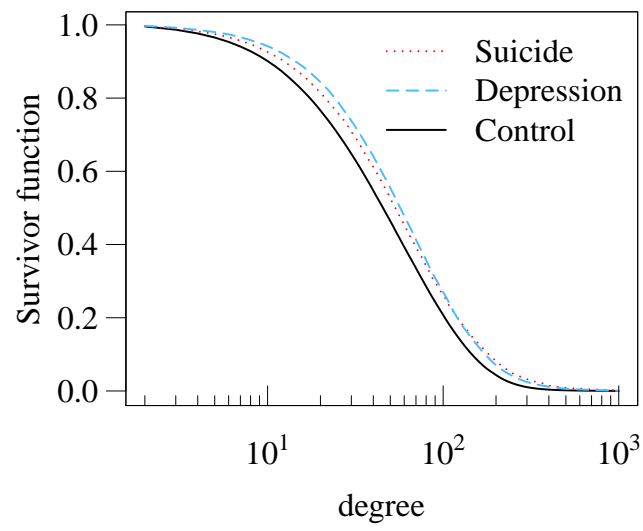


Figure 3

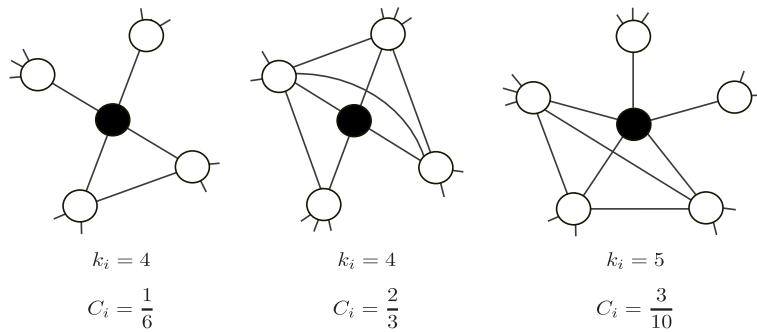


Figure 4

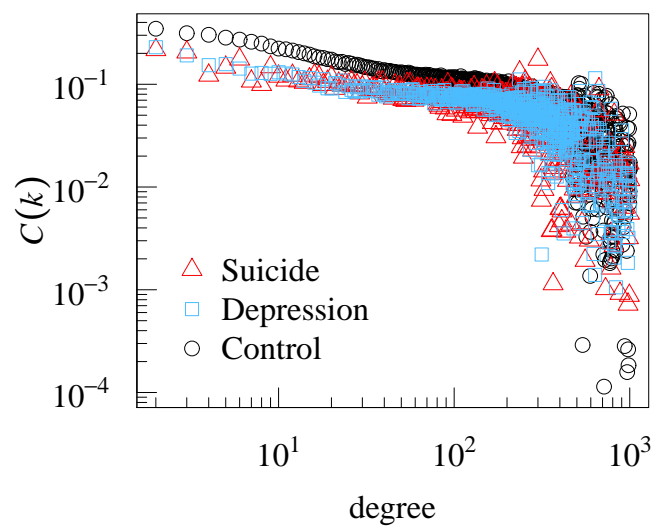


Table 1

ID	Date of creation (day/month/year)	No. users	No. active users	Fraction of active users (%)	No. comments	No. active topics
1	18/01/2008	8367	5985	69.9	741	16
2	21/09/2006	5135	3192	62.9	318	6
3	01/12/2004	3459	1883	53.2	279	12
4	04/02/2008	1445	965	62.4	105	9

Table 2

Variable	Suicide group ($N = 9,990$)		Control group ($N = 228,949$)		p -value
	Mean \pm SD	Range (min,max)	Mean \pm SD	Range (min,max)	
Age	27.4 \pm 10.3	(17, 97)	27.7 \pm 9.2	(14, 96)	0.000652
Community number	283.7 \pm 284.3	(1, 1000)	46.3 \pm 79.4	(1, 1000)	< 0.0001
k_i	82.9 \pm 98.7	(2, 1000)	65.8 \pm 67.6	(2, 1000)	< 0.0001
C_i	0.087 \pm 0.097	(0, 1)	0.150 \pm 0.138	(0, 1)	< 0.0001
Homophily (suicide)	0.0110 \pm 0.0329	(0, 1.000)	0.0012 \pm 0.0080	(0, 0.667)	< 0.0001
Registration period	1235.7 \pm 638.9	(122, 2878)	1333.5 \pm 670.5	(102, 2891)	< 0.0001
Gender (female)	5,786 (57.9%)		126,941 (55.4%)		< 0.0001
No. suicidal communities	1.20 \pm 0.51	(1, 4)	N/A	N/A	N/A
No. login days	28.9 \pm 4.4	(1, 31)	26.9 \pm 6.3	(1, 31)	< 0.0001

Table 3

Variable	OR	CI	<i>p</i> -value	VIF
Age	1.00463	(1.00211, 1.00716)	0.000313	1.091
Gender (female = 1)	0.821	(0.783, 0.861)	< 0.0001	1.028
Community number	1.00733	(1.00720, 1.00747)	< 0.0001	1.197
k_i	0.99790	(0.99758, 0.99821)	< 0.0001	1.156
C_i	0.0093	(0.0069, 0.0126)	< 0.0001	1.081
Homophily (suicide)	2.22×10^{12}	$(0.57 \times 10^{12}, 8.65 \times 10^{12})$	< 0.0001	1.016
Registration period	0.999383	(0.999346, 0.999420)	< 0.0001	1.135

Table 4

Variable 1	Variable 2	Suicide			Depression			Control		
		P	S	K	P	S	K	P	S	K
Age	Gender	-.094	-.137	-.116	-.166	-.174	-.145	-.053	-.026	-.022
Age	Community number	-.045	-.105	-.073	-.089	-.131	-.091	-.032	.023	.015
Age	k_i	-.103	-.224	-.157	-.168	-.268	-.187	-.279	-.385	-.271
Age	C_i	-.048	-.220	-.154	-.092	-.273	-.192	.041	-.152	-.111
Age	Homophily (suicide)	.031	-.037	-.029	N/A	N/A	N/A	-.011	-.090	.074
Age	Homophily (depression)	N/A	N/A	N/A	.166	.121	-.089	-.007	-.083	-.066
Age	Registration period	.159	.356	.259	.203	.364	.266	.278	.460	.337
Gender	Community number	.205	.204	.166	.086	.083	.068	.110	.116	.095
Gender	k_i	.048	.046	.038	.048	.046	.038	.015	.014	.011
Gender	C_i	-.109	-.097	-.080	-.061	-.030	-.024	-.084	-.085	-.069
Gender	Homophily (suicide)	-.007	.031	.028	N/A	N/A	N/A	-.012	-.017	-.017
Gender	Homophily (depression)	N/A	N/A	N/A	-.053	-.021	-.018	.000	.009	.008
Gender	Registration period	-.064	-.061	-.050	-.078	-.079	-.065	.025	.025	.020
Community number	k_i	.348	.338	.231	.375	.360	.248	.375	.372	.258
Community number	C_i	-.231	-.200	-.136	-.201	-.171	-.116	-.376	-.399	-.277
Community number	Homophily (suicide)	-.034	.140	.105	N/A	N/A	N/A	.027	.113	.091
Community number	Homophily (depression)	N/A	N/A	N/A	-.150	.034	.025	.038	.166	.132
Community number	Registration period	.166	.152	.102	.187	.172	.115	.339	.338	.230
k_i	C_i	-.251	-.116	-.085	-.240	-.105	-.074	-.363	-.248	-.175
k_i	Homophily (suicide)	-.175	.174	.107	N/A	N/A	N/A	-.013	.191	.150
k_i	Homophily (depression)	N/A	N/A	N/A	-.210	.076	.029	-.027	.254	.188
k_i	Registration period	.170	.154	.103	.172	.152	.101	.102	.081	.055
C_i	Homophily (suicide)	-.047	-.213	-.162	N/A	N/A	N/A	-.026	-.100	-.080
C_i	Homophily (depression)	N/A	N/A	N/A	-.055	-.243	-.182	-.031	-.145	-.114
C_i	Registration period	-.143	-.112	-.162	-.133	-.099	-.068	-.221	-.249	-.168
Homophily (suicide)	Registration period	-.104	-.059	-.044	N/A	N/A	N/A	-.039	-.031	-.025
Homophily (depression)	Registration period	N/A	N/A	N/A	-.120	-.049	-.036	-.024	.011	.009

Table 5

Variable	OR	CI	<i>p</i> -value	AUC
Age	0.99604	(0.99377, 0.99832)	0.000651	0.515
Gender (female = 1)	1.106	(1.062, 1.152)	< 0.0001	0.512
Community number	1.00728	(1.00716, 1.00741)	< 0.0001	0.867
k_i	1.00259	(1.00237, 1.00280)	< 0.0001	0.549
C_i	0.000581	(0.000428, 0.000789)	< 0.0001	0.690
Homophily (suicide)	1.57×10^{16}	$(0.41 \times 10^{16}, 6.08 \times 10^{16})$	< 0.0001	0.643
Registration period	0.999783	(0.999753, 0.999813)	< 0.0001	0.545

Appendix: Analysis of depressive symptoms

To define depression-related community, we identified the communities satisfying the five criteria as in the case of suicidal community, but with the term suicide in the community name replaced by depression (“utsu” in Japanese). There are ten such communities. We excluded three of them because their names include positive words (let’s overcome; resume one’s place in society, cure; translations by the authors). We define the remaining seven communities, summarized in Tab. A1, to represent depressive symptoms of users. The depression group is the set of active users that belongs to at least one depression-related community listed in Tab. A1. The depression group contains 24410 users. The statistics of the independent variables for the depression group are compared with those for the control group in Figs. 1, 2, 4, and Tab. A2. Each independent variable in the depression and control groups is significantly different in terms of the mean ($p < 0.0001$, Student’s t -test; see Tab. A2) and distribution ($p < 0.0001$, Kolmogorov-Smirnov test).

We applied the multivariate and univariate logistic regressions to identify independent variables that contribute to depressive symptoms (i.e., membership to the depression group). The control group is the same as that used in the main text. The results are shown in Tabs. A3 and A4. The VIF values shown in Tab. A3 and the correlation coefficient values shown in Tab. 4 qualify the use of the multiple logistic regression. The results are qualitatively the same as those for the suicide case.

Table captions

Table A1: Statistics of depression-related communities. For a technical reason, we collected the number of members for communities 1, 2, 3, and 6 on November 2, 2011 and communities 4, 5 and 7 on November 4, 2011.

Table A2: Univariate statistics of independent variables for the depression and control groups. The values for the control group are equal to those shown in Tab. 2 except for those of the homophily variable. The homophily is defined as the fraction of neighbors belonging to the depression group in this table, whereas it is defined as the fraction of neighbors belonging to the suicide group in Tab. 2. The p -value for the gender is based on Chi-square test. The p -values for the other variables are based on Student's t -test. SD: standard deviation.

Table A3: Multivariate logistic regression of depressive symptoms on individual and network variables. OR: odds ratio; CI: 95 % confidence interval; VIF: variance inflation factor.

Table A4: Univariate logistic regression of depressive symptoms on individual and network variables. OR: odds ratio; CI: 95 % confidence interval; AUC: area under the curve.

Table A1

ID	Date of creation (day/month/year)	No. users	No. active users	Fraction of active users (%)	No. comments	No. active topics
1	06/04/2004	15618	8605	54.7	14466	52
2	06/02/2006	13082	9674	72.8	1008	16
3	08/12/2004	4948	2845	56.5	782	17
4	22/04/2006	4606	2907	60.4	221	30
5	28/01/2008	3406	2321	65.0	1350	24
6	09/12/2004	3464	2039	58.2	851	20
7	21/12/2004	2440	1367	54.2	535	5

Table A2

Variable	Depression group ($N = 24,410$)		Control group ($N = 228,949$)		p -value
	Mean \pm SD	Range (min,max)	Mean \pm SD	Range (min,max)	
Age	28.8 \pm 9.4	(16, 97)	27.7 \pm 9.2	(14, 96)	< 0.0001
Community number	249.6 \pm 263.1	(1, 1000)	46.3 \pm 79.4	(1, 1000)	< 0.0001
k_i	81.9 \pm 88.1	(2, 1000)	65.8 \pm 67.6	(2, 1000)	< 0.0001
C_i	0.085 \pm 0.089	(0, 1)	0.150 \pm 0.138	(0, 1)	< 0.0001
Homophily (depression)	0.0196 \pm 0.0501	(0, 1.000)	0.0031 \pm 0.0131	(0, 0.667)	< 0.0001
Registration period	1389.4 \pm 659.2	(122, 2885)	1333.5 \pm 670.5	(102, 2891)	< 0.0001
Gender (female)	16,872 (69.1%)		126,941 (55.4%)		< 0.0001
No. suicidal communities	1.16 \pm 0.47	(1, 6)	N/A	N/A	N/A
No. login days	28.8 \pm 4.4	(1, 31)	26.9 \pm 6.3	(1, 31)	< 0.0001

Table A3

Variable	OR	CI	p -value	VIF
Age	1.0141	(1.0124, 1.0158)	< 0.0001	1.104
Gender (female = 1)	1.532	(1.481, 1.585)	< 0.0001	1.019
Community number	1.00790	(1.00778, 1.00803)	< 0.0001	1.155
k_i	0.99833	(0.99810, 0.99856)	< 0.0001	1.154
C_i	0.0145	(0.0118, 0.0178)	< 0.0001	1.079
Homophily (depression)	1.98×10^{10}	$(0.99 \times 10^{10}, 4.02 \times 10^{10})$	< 0.0001	1.022
Registration period	0.999744	(0.999720, 0.999769)	< 0.0001	1.117

Table A4

Variable	OR	CI	p -value	AUC
Age	1.0110	(1.0097, 1.0123)	< 0.0001	0.551
Gender (female = 1)	1.799	(1.748, 1.850)	< 0.0001	0.568
Community number	1.00826	(1.00814, 1.00837)	< 0.0001	0.860
k_i	1.00258	(1.00243, 1.00274)	< 0.0001	0.566
C_i	0.000415	(0.000338, 0.000509)	< 0.0001	0.692
Homophily (depression)	2.12×10^{12}	$(1.05 \times 10^{12}, 4.28 \times 10^{12})$	< 0.0001	0.658
Registration period	1.000126	(1.000106, 1.000145)	< 0.0001	0.522